Polytomous latent scales for the investigation of the ordering of items

Ligtvoet, R.; van der Ark, L.A.; Bergsma, W.P.; Sijtsma, K.

Published in:
Psychometrika

DOI:
10.1007/s11336-010-9199-8

Publication date:
2011

Document Version
Publisher's PDF, also known as Version of record

PSYCHOMETRIKA—VOL. 76, NO. 2, 200–216
APRIL 2011
DOI: 10.1007/S11336-010-9199-8

POLYTOMOUS LATENT SCALES FOR THE INVESTIGATION OF THE ORDERING OF ITEMS

RUDY LIGTVOET

UNIVERSITY OF AMSTERDAM

L. ANDRIES VAN DER ARK

TILBURG UNIVERSITY

WICHER P. BERGSMA

LONDON SCHOOL OF ECONOMICS

KLAAS SIJTSMA

TILBURG UNIVERSITY

We propose three latent scales within the framework of nonparametric item response theory for polytomously scored items. Latent scales are models that imply an invariant item ordering, meaning that the order of the items is the same for each measurement value on the latent scale. This ordering property may be important in, for example, intelligence testing and person-fit analysis. We derive observable properties of the three latent scales that can each be used to investigate in real data whether the particular model adequately describes the data. We also propose a methodology for analyzing test data in an effort to find support for a latent scale, and we use two real-data examples to illustrate the practical use of this methodology.

Key words: increasingness in transposition, invariant item ordering, latent scales, manifest invariant item ordering, nonparametric IRT models, polytomous IRT models.

1. Introduction

Several applications of tests (or questionnaires) assume that for all individuals to which the instrument is administered item i is more difficult (or less popular) than item j. Generally, the ordering of all the items by their mean scores is assumed to be the same for each person to whom the instrument is administered. This is the assumption of *invariant item ordering* (IIO; Sijtsma & Junker, 1996; Sijtsma & Hemker, 1998). IIO is important, for example, in child intelligence tests when items are administered in the order from easy to difficult, and in personality and attitude measurement when researchers prefer scales that are cumulative or hierarchical.

The purposes of this study are to present and discuss new models for IIO, also known as latent scales (Rosenbaum, 1987a), and methods to investigate the fit of these latent scales to test data. Prior to this, we first define IIO, argue that IIO is important in many test applications, and discuss that the well-known and much-used parametric polytomous IRT models are not suited for investigating IIO in real data.

Requests for reprints should be sent to Rudy Ligtvoet, Faculty of Social and Behavioural Sciences, University of Amsterdam, Nieuwe Prinsengracht 130, 1018 VZ Amsterdam, The Netherlands. E-mail: r.ligtvoet@uva.nl

1.1. Definition of Invariant Item Ordering

Let a test consist of k items, indexed by $i = 1, \ldots, k$. Let random variable X_i denote the item score; X_i has realization x ($x = 0, \ldots, m$), thus assuming equal intervals between adjacent scores. This assumption greatly facilitates the study of IIO and it is also consistent with much psychological research into the way respondents handle rating scale points (Weekers, Brown, & Veldkamp, 2009). The item scores may reflect the degree to which a respondent has solved a cognitive item correctly or endorsed a typical-behavior statement presented in a rating scale item. The latent variable is denoted by θ, and represents the cognitive ability or the personality trait of interest. Finally, $E(X_i|\theta)$ is the conditional expectation of item score X_i, also known as the item response function (IRF; Chang & Mazzeo, 1994). For dichotomously scored items with $x = 0, 1$, we have $E(X_i|\theta) = P(X_i = 1|\theta)$, which is the conditional probability of obtaining a score of 1 on item i.

For polytomously scored items, Sijtsma and Hemker (1998) defined IIO as follows.

Definition. A set of k items with $m + 1$ ordered answer categories per item have IIO if the items can be ordered and numbered accordingly such that

$$E(X_1|\theta) \leq E(X_2|\theta) \leq \cdots \leq E(X_k|\theta), \quad \forall\theta. \tag{1}$$

It may be noted that IIO allows for ties, so that for some values of θ the item ordering is partial.

1.2. Invariant Item Ordering in Test Applications

Several applications of tests use the IIO property but researchers do not always ascertain whether their test has IIO. Instead, they order their items according to the item mean scores in the total group, and assume that this ordering also holds for individuals. Without the support of empirical evidence, simply assuming that the overall item ordering holds for individuals likely means that the researcher makes an aggregation error.

An example of the need for IIO is child intelligence testing. The Wechsler Intelligence Test for Children (Wechsler, 2003) and the Revised Amsterdam Child Intelligence Test (Bleichrodt, Drenth, Zaal, & Resing, 1987) consist of subtests of which the items are administered in order from easy to difficult, based on the proportions of a correct score (the P values). Starting and stopping rules are based on this item ordering. A child in the youngest age group starts with the easiest item and continues until he or she fails a number of consecutive items. Then it stops, as continued failure suggests that the difficulty level has become too high whereas the next items are even more difficult. A child from the next age group skips the first and easiest items because they are too easy for that age group, and then testing again continues until the child fails a number of consecutive items; and so on for the next age groups. These starting and stopping rules are only effective if the ordering of items from easy to difficult is the same for all children—hence, they assume IIO, and empirical research has to support IIO.

Another area in which IIO is relevant is the testing of developmental sequences in cognition that are assumed to be the same for all children. Typical developmental phases are represented in a test by different items that require processes and skills typical of a particular phase but not for others (e.g., Jansen & Van der Maas, 1997). The assumption that individuals develop through the same sequence requires the use of psychometric models that assume a fixed item ordering reflecting this sequence—that is, IIO. If such a model fits the data, it supports the assumption of an invariant developmental sequence; and if it does not fit, it does not support this assumption. Emons, Sijtsma, and Meijer (2007) discussed IIO in person-fit research that uses person response functions. These functions show the probability of a correct answer as item difficulty increases,

and are expected to decrease. Deviations from decreasingness in real data suggest person misfit, and Emons et al. (2007) argued that this interpretation only is useful if IIO holds.

In the typical-behavior domain, researchers often wish their items to have a cumulative or hierarchical structure (e.g., Van Schuur, 2003; Watson, Deary, & Shipley, 2008), because such scales have an unambiguous meaning. Two rating-scale statements expected to have a cumulative or hierarchical structure are 'I do not talk a lot in the company of other people' and 'I prefer not to see people and do things on my own', where the former statement refers to a less intense symptom of introversion, thus inviting higher ratings than the latter. Sijtsma, Meijer, and Van der Ark (2011) argue that a cumulative or hierarchical structure is identical to IIO. If IIO holds, a person with a higher total score has the same symptoms as a person with a lower total score, plus additional symptoms representing higher intensity levels. This hierarchy of symptoms can be inferred from the total score and supports the meaningful interpretation of these total scores, not only as indicators of attribute levels but also as summaries of particular sets of symptoms. Also, IIO implies the same item ordering in interesting subgroups—in this direction, aggregation does not cause errors—and when comparing groups, differences in total-score distributions are easier to interpret.

For dichotomously scored items, the nonparametric Mokken (1971) double monotonicity model and its special case, the Rasch (1960) model, have IIO. Sijtsma and Junker (1996) discussed methods for investigating IIO in Mokken's model, and Glas and Verhelst (1995) discussed methods for investigating goodness-of-fit of the Rasch model. This present study discusses polytomous-item models that have IIO and proposes methods for investigating whether these models are consistent with data.

1.3. Polytomous IRT Models and IIO Research

Sijtsma and Hemker (1998) proved that well-known parametric polytomous IRT models such as the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1990), and the graded response model (Samejima, 1969) do not have IIO. This result means that these well-known models cannot be used to investigate whether a set of items has IIO. For example, a fitting graded response model does not imply IIO for the item set under investigation, and a misfitting graded response model does not imply that the item set does not have IIO.

The reason for the mismatch of popular polytomous IRT models and IIO is the following. Polytomous IRT models define item step response functions (ISRFs) for each score x on item i. For example, the homogeneous case of the graded response model (Samejima, 1969) defines the ISRF as a monotone increasing logistic function with slope parameter α_i and score-category location parameter δ_{ix} $(x_i = 1, \ldots, m)$,

$$C_{ix}(\theta) = P(X_i \geq x|\theta) = \frac{\exp[\alpha_i(\theta - \delta_{ix})]}{1 + \exp[\alpha_i(\theta - \delta_{ix})]}. \tag{2}$$

The ISRFs of an item are related to the item's IRF by

$$E(X_i|\theta) = \sum_{x=1}^{m} C_{ix}(\theta). \tag{3}$$

IIO is defined at the level of IRFs (Equation (1)), and whether or not the ISRFs of a particular model imply IIO, which is defined at the higher aggregation level of IRFs, depends on the precise definition of the ISRFs. Sijtsma and Hemker (1998) proved that Equation (2) does not imply IIO; hence, they proved that the graded response model does not imply IIO and that it is not effective in IIO research.

Sijtsma and Hemker (1998) proved that the rating scale model (Andrich, 1978), which is a special case of the partial credit model, has IIO. Hence, a fitting rating scale model implies

IIO for the item set at hand and could be used in practical IIO research. In addition, Sijtsma and Hemker (1998) defined a restricted version of Muraki's (1990) rating scale version of the graded response model, such that it had IIO and could be used in IIO research. Each of these models is known to be highly restrictive. Hence, we looked for possibly less restrictive models.

For this purpose, we used the following result. Sijtsma and Hemker (1998) defined an order restriction on the ISRFs of the k items in the test that describe the response probability for the same item score x, $C_{ix}(\theta)$, $i = 1, \ldots, k$, such that

$$C_{1x}(\theta) \leq C_{2x}(\theta) \leq \cdots \leq C_{kx}(\theta), \quad \text{for } x = 1, \ldots, m, \ \forall\theta, \tag{4}$$

and showed that Equation (4) implies IIO (Equation (1)) but not the other way around; hence, Equation (4) is a sufficient condition for IIO. Scheiblechner (1995, 2003) discussed *weak item independence*, which resembles Equation (4). Equation (4) does not require a parametric definition of the ISRFs. Given that IIO is a restrictive property and that parametric, polytomous IRT models having IIO are highly restricted versions of more-popular IRT models, we chose to use Equation (4) in a nonparametric approach, which is less restrictive and is discussed in this study.

In the next sections, we discuss three classes of polytomous-item IRT models, and implement inequality constraints comparable to those in Equation (4) in each of the three classes. We prove that the resulting three classes of models are hierarchically related, and that all three have IIO. We derive observable consequences, propose different methods for investigating these consequences in real data, and illustrate the methods in two real-data examples.

2. Three Classes of Polytomous IRT Models

Polytomous IRT models are commonly divided into the *cumulative probability models*, *continuation ratio models*, and *adjacent category models* (Agresti, 1990; Hemker, Van der Ark, & Sijtsma, 2001; Mellenbergh, 1995; Molenaar, 1983). Each class assumes unidimensionality; that is, the k items in the test share one unidimensional latent variable, and local independence; that is, for a k-dimensional vector of item scores $\mathbf{X} = \mathbf{x}$,

$$P(\mathbf{X} = \mathbf{x}|\theta) = \prod_{i=1}^{k} P(X_i = x|\theta). \tag{5}$$

An item having $m + 1$ ordered answer categories has m *item steps*, which have to be passed in going from category 0 to category m (Molenaar, 1983). The probability of passing the item step conditional on θ is the ISRF; for example, see Equation (2). The three classes of IRT models differ in their definition of the ISRF, and models within classes place different restrictions on their class-specific ISRF.

Cumulative probability models (CPMs) define ISRFs as

$$C_{ix}(\theta) = P(X_i \geq x|\theta) = \sum_{u=x}^{m} P(X_i = u|\theta), \tag{6}$$

for $x = 1, \ldots, m$, and $C_{i0}(\theta) = 1$ for $x < 1$, and $C_{ix}(\theta) = 0$ for $x > m$. This ISRF definition implies that the ISRFs of the same item cannot intersect (Mellenbergh, 1995). Examples of CPMs are the homogeneous case of the graded response model (Samejima 1969, 1997; Equation (2)), and the nonparametric graded response model (Hemker, Sijtsma, Molenaar, & Junker, 1997; also, see Molenaar, 1997). These models assume that the ISRF defined by $C_{ix}(\theta)$ (Equation (6)) increases monotonically (i.e., the monotonicity assumption). Van Engelenburg (1997,

Chapters 2, 3) argued that CPMs are particulary suited for modeling item scores that result from a global assessment task as with rating scales.

Continuation ratio models (CRMs) define ISRFs as

$$M_{ix}(\theta) = P(X_i \geq x | X_i \geq x-1;\theta) = \frac{\sum_{u=x}^{m} P(X_i = u|\theta)}{\sum_{v=x-1}^{m} P(X_i = v|\theta)}, \tag{7}$$

for $x = 1, \ldots, m$, and $M_{i0}(\theta) = 1$ for $x < 1$, and $M_{ix}(\theta) = 0$ for $x > m$. Examples of CRMs are the sequential Rasch model (Tutz, 1990), and the nonparametric sequential model (Hemker et al., 2001). These models assume monotonicity for $M_{ix}(\theta)$ (Equation (7)). Items typically suited for CRM analysis consist of m subtasks that have to be executed in a fixed order such that failing a subtask implies failing the next subtasks, and the item score reflects that the first x subtasks have been successfully executed (Van Engelenburg, 1997, Chapters 2, 3; Hemker et al., 2001).

Adjacent category models (ACMs) define ISRFs as

$$\begin{aligned} A_{ix}(\theta) &= P(X_i = x | X_i = x \vee X_i = x-1;\theta) \\ &= \frac{P(X_i = x|\theta)}{P(X_i = x-1|\theta) + P(X_i = x|\theta)}, \end{aligned} \tag{8}$$

for $x = 1, \ldots, m$, and $A_{i0}(\theta) = 1$ for $x < 1$, and $A_{ix}(\theta) = 0$ for $x > m$. Examples of ACMs include the rating scale model (Andrich, 1978), the partial credit model (Masters, 1982), the generalized partial credit model (Muraki, 1992), and the nonparametric partial credit model (Hemker et al., 1997). These models assume monotonicity for $A_{ix}(\theta)$ (Equation (8)). Van Engelenburg (1997, p. 38) suggested that ACMs are best suited for analyzing item scores that result from tasks that consist of x subtasks, which may be solved in an arbitrary order. An item score of x means that any x subtasks were solved correctly.

Van der Ark, Hemker and Sijtsma (2002) showed that the mathematically most general representatives of each of the three classes, which are the nonparametric graded response model (CPM class), the nonparametric sequential model (CRM class), and the nonparametric partial credit model (ACM class), have a hierarchical relationship; that is, using obvious acronyms,

$$\text{np-ACM} \quad \Rightarrow \quad \text{np-CRM} \quad \Rightarrow \quad \text{np-CPM}.$$

Thus, $A_{ix}(\theta)$ (ACM class) provides the strongest form of monotonicity, and $C_{ix}(\theta)$ (CPM class) the weakest. For dichotomously scored items, the three classes coincide, such that $C_{ix}(\theta) = M_{ix}(\theta) = A_{ix}(\theta) = P(X_i = 1|\theta)$.

We generalize Equation (4) for the CPM class to the typical ISRFs of the CRM and ACM classes. This results in three nonparametric IRT models that imply IIO. We prove that the three nonparametric polytomous IRT models have a hierarchical relationship and derive observable consequences, which are used to investigate in real data whether a set of k items has IIO.

3. Latent Scales for Polytomous Items

For dichotomously scored items, Rosenbaum (1987a) defined a latent scale as a model for which local independence (i.e., Equation (5)) holds and in each item pair (i, j) $(i < j)$ item i is uniformly more difficult than item j (Rosenbaum, 1987b), so that

$$P(X_i = 1|\theta) \leq P(X_j = 1|\theta), \quad \forall\theta.$$

We generalize the concept of a latent scale to the three classes of polytomous IRT models. Using the acronym LS for latent scale, the resulting models are denoted LS-CPM, LS-CRM, and LS-ACM.

Definition. We assume local independence for the k polytomously scored items in the test. For scores $x = 1, \ldots, m$ on items i and j $(i < j)$, an LS-CPM is defined as

$$C_{ix}(\theta) \leq C_{jx}(\theta), \quad \forall\theta \tag{9}$$

(equivalent to Equation (4)); an LS-CRM as

$$M_{ix}(\theta) \leq M_{jx}(\theta), \quad \forall\theta; \tag{10}$$

and an LS-ACM as

$$A_{ix}(\theta) \leq A_{jx}(\theta), \quad \forall\theta. \tag{11}$$

Equations (9), (10), and (11) do not restrict the ordering of the ISRFs corresponding to different score categories. Equation (9) is equivalent to Equation (4) and thus implies IIO. The latent scales do not assume monotonicity. We next provide and prove a theorem on a hierarchical relationship between LS-ACM, LS-CRM, and LS-CPM.

Theorem 1. *The three latent-scale IRT models for polytomously scored items, the LS-ACM, the LS-CRM, and the LS-CPM have a hierarchical relationship. The least restrictive of these models, the LS-CPM, implies IIO. These relationships are represented in the following scheme of logical implications*:

$$\textit{LS-ACM} \quad \Rightarrow \quad \textit{LS-CRM} \quad \Rightarrow \quad \textit{LS-CPM} \quad \Rightarrow \quad \textit{IIO}.$$

We prove three lemmas, which together prove Theorem 1.

Lemma 1. *The LS-ACM implies the LS-CRM.*

Proof: First note that, for $z > x$,

$$\begin{aligned}
A_{iy}(\theta) \leq A_{jy}(\theta) \quad &\Leftrightarrow \quad \frac{1 - A_{iy}(\theta)}{A_{iy}(\theta)} \geq \frac{1 - A_{jy}(\theta)}{A_{jy}(\theta)} \\
&\Leftrightarrow \quad \frac{P(X_i = y-1|\theta)}{P(X_i = y|\theta)} \geq \frac{P(X_j = y-1|\theta)}{P(X_j = y|\theta)} \\
&\Rightarrow \quad \prod_{y=x+1}^{z} \frac{P(X_i = y-1|\theta)}{P(X_i = y|\theta)} \geq \prod_{y=x+1}^{z} \frac{P(X_j = y-1|\theta)}{P(X_j = y|\theta)} \\
&\Leftrightarrow \quad \frac{P(X_i = x|\theta)}{P(X_i = z|\theta)} \geq \frac{P(X_j = x|\theta)}{P(X_j = z|\theta)} \\
&\Leftrightarrow \quad \frac{P(X_i = z|\theta)}{P(X_i = x|\theta)} \leq \frac{P(X_j = z|\theta)}{P(X_j = x|\theta)}.
\end{aligned} \tag{12}$$

Thus, we have shown that LS-ACM (Equation (11)) implies Equation (12). Summing both sides of Equation (12) over $z = x+1, x+2, \ldots, m$ gives

$$\frac{P(X_i > x|\theta)}{P(X_i = x|\theta)} \leq \frac{P(X_j > x|\theta)}{P(X_j = x|\theta)},$$

which implies

$$\frac{P(X_i = x|\theta)}{P(X_i > x|\theta)} \geq \frac{P(X_j = x|\theta)}{P(X_j > x|\theta)},$$

and so

$$\begin{aligned}\frac{P(X_i \geq x|\theta)}{P(X_i > x|\theta)} &= \frac{P(X_i = x|\theta) + P(X_i > x|\theta)}{P(X_i > x|\theta)} \\ &= \frac{P(X_i = x|\theta)}{P(X_i > x|\theta)} + 1 \geq \frac{P(X_j = x|\theta)}{P(X_j > x|\theta)} + 1 = \frac{P(X_j \geq x|\theta)}{P(X_j > x|\theta)}. \end{aligned} \tag{13}$$

The left- and right-hand sides of Equation (13) are the reciprocals of $M_{i,x+1}(\theta)$ and $M_{j,x+1}(\theta)$, respectively, so we have shown that LS-ACM implies

$$M_{ix}(\theta) \leq M_{jx}(\theta)$$

for all x, θ and $i < j$; that is, that LS-CRM holds. □

The following example shows that the reverse relationship between the two latent scales does not hold; that is, the LS-CRM does not imply the LS-ACM. For trichotomously scored items i and j, for some arbitrary value θ_0 let the item scores (0, 1, 2) have probabilities $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ (item i) and $(\frac{1}{3}, \frac{1}{12}, \frac{7}{12})$ (item j). It may be verified that $M_{i1} = M_{i2} = \frac{1}{2}$, $M_{j1} = \frac{2}{3}$, and $M_{j2} = \frac{7}{8}$, and further that for $x = 1, 2$ it holds that $M_{ix} < M_{jx}$. Additional computations show that $A_{i1} = \frac{1}{3}$, $A_{i2} = \frac{1}{2}$, $A_{j1} = \frac{1}{5}$, and $A_{j2} = \frac{7}{8}$. Because $A_{i1} > A_{j1}$ contradicts Equation (11), the LS-ACM does not hold.

Lemma 2. *The LS-CRM implies the LS-CPM.*

Proof: We assume that the LS-CRM holds; that is, Equation (10) holds for all x and all θ. It may be noted that

$$\begin{aligned} C_{ix}(\theta) &= \frac{P(X_i \geq x|\theta)}{P(X_i \geq 0|\theta)} \\ &= \frac{P(X_i \geq 1|\theta)}{P(X_i \geq 0|\theta)} \times \frac{P(X_i \geq 2|\theta)}{P(X_i \geq 1|\theta)} \times \cdots \times \frac{P(X_i \geq x|\theta)}{P(X_i \geq x-1|\theta)} \\ &= \prod_{u=1}^{x} M_{iu}(\theta). \end{aligned} \tag{14}$$

Because $M_{ix}(\theta) \leq M_{jx}(\theta)$ for all x and all θ, it follows from Equation (14) that

$$C_{ix}(\theta) \leq C_{jx}(\theta);$$

that is, that LS-CPM holds. □

The following example shows that the reverse of the implication does not hold; that is, the LS-CPM does not imply the LS-CRM. For trichotomously scored items i and j, for some arbitrary value θ_0, let the item scores (0, 1, 2) have probabilities $(\frac{1}{2}, \frac{1}{4}, \frac{1}{4})$ (item i) and $(\frac{1}{3}, \frac{9}{24}, \frac{7}{24})$ (item j). It may be verified that $C_{i1} = \frac{1}{2}$, $C_{i2} = \frac{1}{4}$, $C_{j1} = \frac{2}{3}$, and $C_{j2} = \frac{7}{24}$. Next, it can be verified that $C_{ix} < C_{jx}$ for $x = 1, 2$. Finally, we find that $M_{i1} = M_{i2} = \frac{1}{2}$, $M_{j1} = \frac{2}{3}$, and $M_{j2} = \frac{7}{16}$. Because $M_{i2} > M_{j2}$ contradicts Equation (10), the LS-CRM does not hold.

Lemma 3. *The LS-CPM implies IIO.*

Proof: See Sijtsma and Hemker (1998), who show that Equation (9) is a sufficient (but not a necessary) condition for IIO. □

The three latent scales provide different definitions of agreement among the respondents with respect to the ordering of the items on latent variable θ (for related work, see Douglas, Fienberg, Lee, Sampson, & Whitaker, 1991). A fourth latent scale may be defined by the combination of local independence and IIO. Theorem 1 shows that the four latent-scale definitions become progressively weaker, going from the LS-ACM via the LS-CRM and the LS-CPM to the latent scale defined by IIO and local independence. Thus far, in psychometrics item orderings have been defined in terms of expected item scores such as IIO in Equation (1). IIO is the weakest form of agreement among the respondents with respect to the ordering of the items on latent variable θ. Given their relationships to particular task structures (Van Engelenburg, 1997), the other latent scales are also plausible ways of defining this agreement.

4. Manifest Properties of Latent Scales

In this section, we derive three observable consequences or manifest properties from the latent scales. In particular, we prove that the LS-ACM implies the *increasingness in transposition* (IT) property (Theorem 3); the LS-CPM implies the *manifest scale cumulative probability model* (MS-CPM) property (Theorem 2); and IIO implies the *manifest invariant item ordering* (MIIO) property (Corollary). Latent scales and observable properties are proved to be related as follows:

$$\begin{array}{ccccccc} \text{LS-ACM} & \Rightarrow & \text{LS-CRM} & \Rightarrow & \text{LS-CPM} & \Rightarrow & \text{IIO} \\ \Downarrow & & & & \Downarrow & & \Downarrow \\ \text{IT} & & & & \text{MS-CPM} & & \text{MIIO} \end{array}$$

The manifest properties can be used as a basis for investigating whether support can be found in the data for a particular latent scale. To this end, we discuss the IT method, the MS-CPM method, and the MIIO method in the next section.

4.1. Manifest Scales

Let Y be a manifest variable with realization y that is independent of item scores X_i and X_j given θ. For example, Y may be a function of the $k-2$ items in the test without the items i and j, the sum score obtained on a different test, or an indicator of group membership. Replacing the latent variable θ in the latent scales defined in Equations (9), (10), and (11) by the manifest variable Y yields their manifest scale (MS) analogues. Thus, for $i < j$ and for $x = 1, \ldots, m$, an MS-CPM is defined as $C_{ix}(Y) \leq C_{jx}(Y)$ for all values of Y (cf. Equation (9)), an MS-CRM is defined as $M_{ix}(Y) \leq M_{jx}(Y)$ for all values of Y (cf. Equation (10)), and an MS-ACM is defined as $A_{ix}(Y) \leq A_{jx}(Y)$ for all values of Y (cf. Equation (11)). Similarly, for $i < j$ an MIIO is defined as $E(X_i|Y = y) \leq E(X_j|Y = y)$ for all y (cf. Equation (1)).

In Theorem 2, we prove that the LS-CPM implies the MS-CPM, and the Corollary shows that IIO implies an MIIO. The LS-ACM and the LS-CRM do not imply their manifest analogues. Thus, if in empirical data analysis the MS-CPM and MIIO are satisfied, some support is found for the theoretical LS-CPM and IIO, respectively, but if the MS-ACM and the MS-CRM are satisfied, this cannot be taken as support for their latent scale analogues. Because fitting MS-ACMs and MS-CRMs do not support latent scales, they are not further pursued here.

Theorem 2. *The LS-CPM implies the MS-CPM.*

Proof: Let $F(\theta)$ be the cumulative distribution function of θ. Multiplying both sides of Equation (9) by $P(Y = y|\theta)$ and integrating over θ yields

$$\begin{aligned} & C_{ix}(\theta) \leq C_{jx}(\theta), \quad \forall\theta && (15)\\ \Leftrightarrow \quad & P(X_i \geq x|\theta) \leq P(X_j \geq x|\theta), \quad \forall\theta \\ \Rightarrow \quad & \int_\theta P(X_i \geq x|\theta)P(Y = y|\theta)\,dF(\theta) \leq \int_\theta P(X_j \geq x|\theta)P(Y = y|\theta)\,dF(\theta). && (16) \end{aligned}$$

Because Y is conditionally independent of X_i and X_j, Equation (16) is equivalent to

$$\begin{aligned} & \int_\theta P(X_i \geq x, Y = y|\theta)\,dF(\theta) \leq \int_\theta P(X_j \geq x, Y = y|\theta)\,dF(\theta) \\ \Leftrightarrow \quad & P(X_i \geq x, Y = y) \leq P(X_j \geq x, Y = y) \\ \Leftrightarrow \quad & P(X_i \geq x|Y = y) \leq P(X_j \geq x|Y = y). && (17) \end{aligned}$$

The proof holds for all x and y, and all $i < j$. □

The reverse is not true, MS-CPM does not imply the LS-CPM. Also, IIO does not imply MS-CPM, which means that a violation of MS-CPM does not disprove IIO. For reasons of limited space, we do not provide counter examples but refer to Ligtvoet, Van der Ark, Bergsma, and Sijtsma (2010b). For an appropriate choice of variable Y, in real data it can be investigated whether the MS-CPM property (Equation (17)) is satisfied. Variable Y should be closely related to θ, and a likely choice may be the sum score on a subset of items from the test that also measures θ. The next section discusses how the MS-CPM method based on Equation (17) may be used in the analysis of real data.

Corollary. *IIO implies MIIO.*

Proof: Theorem 2 states for $x = 1, \ldots, m$, for $i < j$, and all θ that

$$P(X_i \geq x|\theta) \geq P(X_j \geq x|\theta) \quad \Rightarrow \quad P(X_i \geq x|Y = y) \geq P(X_j \geq x|Y = y)$$

for all y. This result can also be shown to hold for sums of cumulative response probabilities. For $x = 1, \ldots, m$, for $i < j$, and all θ

$$\begin{aligned} & \sum_{x=1}^{m} P(X_i \geq x|\theta) \geq \sum_{x=1}^{m} P(X_j \geq x|\theta) \\ \Rightarrow \quad & \sum_{x=1}^{m} P(X_i \geq x|Y = y) \geq \sum_{x=1}^{m} P(X_j \geq x|Y = y) \end{aligned}$$

for all y. This implication is equivalent to

$$E(X_i|\theta) \geq E(X_j|\theta) \quad \Rightarrow \quad E(X_i|Y = y) \geq E(X_j|Y = y); \tag{18}$$

also see Shaked and Shantikumar (1994, p. 4). □

The reverse is not true, MIIO does not imply IIO. For an appropriate choice of variable Y, the MIIO property (i.e., the right-hand side of Equation (18)) can be investigated in real data so as to collect support in favor of IIO. The corresponding MIIO method is discussed in the next section.

Because IIO implies MIIO, by implication the previous three latent scales in the ordered series also imply MIIO. In the same vein, because the LS-CPM implies the MS-CPM, the preceding and most restrictive latent scales, the LS-ACM and the LS-CRM, also imply the MS-CPM.

4.2. Increasingness in Transposition

Rosenbaum (1987a) used the manifest IT property (Hollander, Proschan, & Sethuraman, 1977) to investigate whether a set of dichotomously scored items forms a latent scale. We adapt the results presented by Rosenbaum (1987a) to investigate whether a set of polytomously scored items constitute a latent scale (Equations (9), (10), and (11)). First, we introduce some notation.

The set of items and their indices $\mathcal{T}$ is divided into two subsets $(\mathcal{S}, \mathcal{R})$. Subset $\mathcal{S}$ contains at least two items, and subset $\mathcal{R}$ contains the remaining items. The realization of the scores on the items in $\mathcal{S}$ are collected in item-score vector $\mathbf{x}_\mathcal{S}$, and the scores on the items in $\mathcal{R}$ in item-score vector $\mathbf{x}_\mathcal{R}$. We define item difficulty as the expected score on an item across the distribution of θ, denoted $F(\theta)$: that is, $E(X_i) = \int E(X_i|\theta)\, dF(\theta)$, for $i = 1, \ldots, k$. Let i and j be two items from $\mathcal{S}$, and let $i < j$ denote that item i is at least as difficult as item j; that is, $E(X_i) \leq E(X_j)$. Then, $x_i > x_j$ means that the score on the more difficult item i is higher than the score on the easier item j. Furthermore, let $h(\mathbf{X}_\mathcal{R})$ be a function of the scores on the items in $\mathcal{R}$. For example, $h(\mathbf{X}_\mathcal{R})$ may be the sum score on the items in $\mathcal{R}$, or it may be a single item score.

Vector $\mathbf{x}'_\mathcal{S}$ is defined as a *transposition* of vector $\mathbf{x}_\mathcal{S}$, if one or more reversals of two scores in vector $\mathbf{x}_\mathcal{S}$ produce vector $\mathbf{x}'_\mathcal{S}$ (Hollander et al., 1977). For example, $\mathbf{x}'_\mathcal{S} = (1, 1, 0, 2)$ is a transposition of $\mathbf{x}_\mathcal{S} = (1, 2, 0, 1)$, because the reversal of x_2 and x_4 in $\mathbf{x}_\mathcal{S}$ produces $\mathbf{x}'_\mathcal{S}$. Two reversals are needed to go from $\mathbf{x}_\mathcal{S}$ to $\mathbf{x}''_\mathcal{S} = (0, 1, 1, 2)$. Finally, $\mathbf{x}''_\mathcal{S} = (1, 2, 1, 2)$ and $\mathbf{x}_\mathcal{S}$ are not transpositions of one another.

Next, we consider two vectors $\mathbf{x}_\mathcal{S}$ and $\mathbf{x}'_\mathcal{S}$, which are transpositions of one another, and define the *partial order* '$\prec$' on these vectors. A partial order $\mathbf{x}_\mathcal{S} \prec \mathbf{x}'_\mathcal{S}$ means that $\mathbf{x}_\mathcal{S}$ produces $\mathbf{x}'_\mathcal{S}$ when interchanging item scores in $\mathbf{x}_\mathcal{S}$ implies that higher item scores z are moved to the right while lower item scores y are moved to the left. In the previous example, $\mathbf{x}_\mathcal{S}$ produced $\mathbf{x}'_\mathcal{S}$ when the higher score $x_2 = 2$ was interchanged with the lower score $x_4 = 1$. What happens is that, given the item ordering $E(X_1) \leq E(X_2) \leq E(X_3) \leq E(X_4)$, the ordering of item scores in the resulting vector $\mathbf{x}'_\mathcal{S}$ better matches the item ordering by difficulty than in the original vector $\mathbf{x}_\mathcal{S}$.

Let $P[\mathbf{x}_\mathcal{S}|h(\mathbf{X}_\mathcal{R})]$ be the probability of item-score vector $\mathbf{x}_\mathcal{S}$ conditional on score function $h(\mathbf{X}_\mathcal{R})$. Under some IRT models, such probabilities can be ordered in $\mathbf{X}_\mathcal{S}$ (i.e., for different vectors $\mathbf{x}_\mathcal{S}$) provided the item-score vectors are partially ordered. More specifically, conditional on function $h(\mathbf{X}_\mathcal{R})$, the probabilities of two vectors $\mathbf{x}_\mathcal{S}$ and $\mathbf{x}'_\mathcal{S}$, which are partially ordered by $\mathbf{x}_\mathcal{S} \prec \mathbf{x}'_\mathcal{S}$, are ordered such that $P[\mathbf{x}_\mathcal{S}|h(\mathbf{X}_\mathcal{R})] \leq P[\mathbf{x}'_\mathcal{S}|h(\mathbf{X}_\mathcal{R})]$. When such an ordering is possible, the probabilities are *increasing in transposition* in $\mathbf{X}_\mathcal{S}$. Suppose, the partially ordered vectors $\mathbf{x}_\mathcal{S}$ and $\mathbf{x}'_\mathcal{S}$ differ with respect to two or more transpositions; then, successive transpositions step-by-step move higher scores from $\mathbf{x}_\mathcal{S}$ to the right until $\mathbf{x}'_\mathcal{S}$ is obtained. Vectors $\mathbf{x}_\mathcal{S}$ and $\mathbf{x}'_\mathcal{S}$ and the vectors obtained in each step moving from $\mathbf{x}_\mathcal{S}$ to $\mathbf{x}'_\mathcal{S}$ are collected in a set denoted $\mathcal{V}$. It may be noted that set $\mathcal{V}$ contains only those vector permutations that are partially ordered. Then, the formal definition of functions that are IT in $\mathbf{X}_\mathcal{S}$ is the following.

Definition. $P(.)$ is IT in $\mathbf{X}_{\mathcal{S}}$ for function $h(.)$ if for all $\{\mathbf{x}_{\mathcal{S}}, \mathbf{x}'_{\mathcal{S}}\} \in \mathcal{V}$, which have a partial ordering $\mathbf{x}_{\mathcal{S}} \prec \mathbf{x}'_{\mathcal{S}}$, we have

$$P\big[\mathbf{x}_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})\big] \leq P\big[\mathbf{x}'_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})\big].$$

As an example, for the sake of simplicity we assume that $\mathcal{R} = \emptyset$. Thus, $\mathcal{S} = \mathcal{T}$, so that $P[\mathbf{x}_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})] = P(\mathbf{x}_{\mathcal{S}})$. Now, because the vectors $(2, 1, 1, 0)$ and $(0, 1, 1, 2)$ are partially ordered, the IT property implies that $P(2, 1, 1, 0) \leq P(0, 1, 1, 2)$.

Theorem 3. *The LS-ACM implies IT.*

Proof: The point of departure is Equation (12), which holds under the LS-ACM. For $0 \leq y < z \leq m$ and $i < j$, Equation (12) is equivalent to

$$\frac{P(X_i = z|\theta)P(X_j = y|\theta)}{P(X_i = y|\theta)P(X_j = z|\theta)} \leq 1. \tag{19}$$

For dichotomously scored items with $y = 0$ and $z = 1$, Rosenbaum (1987a, Theorem 1) showed that Equation (19) implies IT. We extend Rosenbaum's proof to polytomous items.

Let k_s be the number of items in subset $\mathcal{S}$, and let $\mathcal{S} \setminus \{i, j\}$ denote the subset of $k_s - 2$ items that remain in $\mathcal{S}$ after items i and j have been excluded. For subset $\mathcal{S}$ including items i and j (i.e., $\{i, j\} \in \mathcal{S}$), Equation (19) is equivalent to

$$\frac{P(X_i = z|\theta)P(X_j = y|\theta)}{P(X_i = y|\theta)P(X_j = z|\theta)} \prod_{u \in \mathcal{S}\setminus\{i,j\}} \frac{P(X_u = x_u|\theta)}{P(X_u = x_u|\theta)} \leq 1. \tag{20}$$

Because of local independence (Equation (5)), Equation (20) can be written as

$$\frac{P(X_1 = x_1, \ldots, X_i = z, \ldots, X_j = y, \ldots, X_{k_s} = x_{k_s}|\theta)}{P(X_1 = x_1, \ldots, X_i = y, \ldots, X_j = z, \ldots, X_{k_s} = x_{k_s}|\theta)} \leq 1. \tag{21}$$

The item-score vector in the numerator is denoted by $\mathbf{x}_{\mathcal{S}}$ and the item-score vector in the denominator by $\mathbf{x}'_{\mathcal{S}}$. It may be noted that $\mathbf{x}_{\mathcal{S}}$ and $\mathbf{x}'_{\mathcal{S}}$ are partially ordered, $\mathbf{x}_{\mathcal{S}} \prec \mathbf{x}'_{\mathcal{S}}$. We rewrite Equation (21) as

$$\frac{P(\mathbf{x}_{\mathcal{S}}|\theta)}{P(\mathbf{x}_{\mathcal{S}'}|\theta)} \leq 1. \tag{22}$$

Hollander et al. (1977, Theorem 3.2) show that Equation (22) implies

$$\int \frac{P(\mathbf{x}_{\mathcal{S}}|\theta)}{P(\mathbf{x}'_{\mathcal{S}}|\theta)}\, dF(\theta) \leq 1. \tag{23}$$

Finally, Equation (23) implies the manifest IT property,

$$\frac{P[\mathbf{x}_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})]}{P[\mathbf{x}'_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})]} \leq 1 \quad \Leftrightarrow \quad P\big[\mathbf{x}_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})\big] \leq P\big[\mathbf{x}'_{\mathcal{S}}|h(\mathbf{X}_{\mathcal{R}})\big] \tag{24}$$

(cf. Rosenbaum, 1987a, Theorem 1). □

The other latent scales (LS-CRM, LS-CPM) and IIO do not imply IT. Ligtvoet et al. (2010b), provide counter examples. It may be noted that Equation (24) holds for any conditioning variable

Y that is independent of the items in $\mathcal{S}$, but that we used $h(\mathbf{X}_{\mathcal{R}})$ to stay close to previous work (Rosenbaum, 1987a; Sijtsma & Junker, 1996). The proof may be extended to partially ordered vectors $\mathbf{x}_{\mathcal{S}}$ and $\mathbf{x}'_{\mathcal{S}}$ that differ with respect to two or more transpositions following a step-by-step permutation of $\mathbf{x}_{\mathcal{S}}$ into $\mathbf{x}'_{\mathcal{S}}$ by successive transpositions that move higher scores to the right, and applying Equation (21) successively. In the next section, we discuss how the IT method based on the IT property can be used in the analysis of real data to collect support in favor of the LS-ACM.

5. Methods for Data Analysis

For realistic numbers of items, the investigation of the MIIO, MS-CPM, and IT properties produces multiple results, which have to be combined for each property to decide whether that property holds in the data and, hence, provides support for a particular latent scale. Ligtvoet, Van der Ark, Te Marvelde, and Sijtsma (2010a) proposed a method for dealing with multiple results when testing the MIIO property. Here, we adapt this method to the MS-CPM and IT properties, but first we explain the MIIO method (see Ligtvoet et al., 2010a, for details).

For each item pair (i, j) $(i < j)$, it is investigated whether it violates MIIO (Equation (18)). This produces $\frac{1}{2} \times k \times (k-1)$ Boolean outcomes on violation of MIIO. The statistical testing procedure for one item pair (i, j) is as follows. Variable Y in Equation (18) is replaced by rest score $R_{(ij)} = \sum_{g \neq i,j} X_g$, so that MIIO is investigated by checking whether $E(X_i|R_{(ij)} = r) \leq E(X_j|R_{(ij)} = r)$ for $r = 0, \ldots, k-2$. If the sample means are reversely ordered (i.e., $\overline{X}_i|R_{(ij)} = r > \overline{X}_j|R_{(ij)} = r$), a one-sided t-test is used for deciding whether the violation is significant. To avoid testing violations that are too small on a scale ranging from 0 to m to be of practical interest, violations smaller than $m \times 0.03$ are ignored. Adjacent rest-score groups $r, r+1, \ldots$ containing few observations may be joined to gain statistical power (Molenaar & Sijtsma, 2000, p. 67). If one or more t-tests of violations in excess of $m \times 0.03$ are significant, the item pair violates MIIO.

If MIIO does not hold for each of the $\frac{1}{2} \times k \times (k-1)$ item pairs, items are removed one-by-one until a subset remains for which MIIO holds (Ligtvoet et al., 2010a). A backward item-selection procedure reaches this goal while removing as few items as possible. This is done in the first step by counting for each item how many of the $k-1$ item pairs in which the item is involved violate MIIO significantly according to the t-test procedure. The item with the highest count is removed first; for the remaining $k-1$ items the counts are redone without the item that was removed, and if there are item pairs violating MIIO, the item having the highest count is removed; and this procedure is repeated until there are no item pairs left that violate MIIO. If two or more items have the highest count, then the item that has the lowest scalability value is removed (Ligtvoet et al., 2010a). The same rest score based on $k-2$ items is used throughout so as to minimize the risk of chance capitalization. We adapt this strategy to the MS-CPM (Equation (17)) and IT (Equation (24)) properties, thus producing the MS-CPM and IT methods.

For the MS-CPM property, let $P(X_i \geq x|Y = y)$ and $P(X_j \geq x|Y = y)$ in Equation (17) be denoted *pair of manifest ISRFs* (i, j, x). For each pair of manifest ISRFs (rather than for each item pair) it is investigated whether the pair violates the MS-CPM property, which produces $\frac{1}{2} \times k \times (k-1) \times m$ Boolean outcomes. The testing procedure for one pair of manifest ISRFs (i, j, x) is as follows. Just as for MIIO, variable Y in Equation (17) is replaced by rest score $R_{(ij)}$. Hence, the MS-CPM property is investigated by checking whether $P(X_i \geq x|R_{(ij)} = r) \leq P(X_j \geq x|R_{(ij)} = r)$ for all r. If the sample fractions are reversely ordered (i.e., $\hat{P}(X_i \geq x|R_{(ij)} = r) > \hat{P}(Xj \geq x|R_{(ij)} = r)$, a z-test (Molenaar & Sijtsma, 2000, p. 78) is used to decide whether the violation is significant . Following recommendations by Molenaar and Sijtsma (2000, pp. 67–70), violations smaller than 0.03 are ignored.

For the IT property (Equation (24)), the method is adapted as follows. We consider item pairs, so that item-score vectors $\mathbf{x}_{\mathcal{S}}$ and $\mathbf{x}_{\mathcal{S}'}$ in Equation (24) are reduced to two elements: $\mathbf{x}_{\mathcal{S}} = (u, v)$ and $\mathbf{x}_{\mathcal{S}'} = (v, u)$ $(u = 0, \ldots, m-1; v = u+1, \ldots, m)$. Consistent with MIIO and MS-CPM, function $h(\mathbf{X}_{\mathcal{R}}) = R_{(ij)}$. Let $P(\mathbf{X}_{\mathcal{S}} = \mathbf{x}_{\mathcal{S}} | h(\mathbf{X}_{\mathcal{R}})) = P(X_i = u, X_j = v | R_{(ij)})$ and $P(\mathbf{X}_{\mathcal{S}'} = \mathbf{x}_{\mathcal{S}'} | h(\mathbf{X}_{\mathcal{R}})) = P(X_i = v, X_j = u | R_{(ij)})$ be the *pair of bivariate conditional probabilities* (i, j, u, v). For each pair of bivariate conditional probabilities, it is investigated whether the pair violates the IT property. This produces $\frac{1}{2} \times k \times (k-1) \times \frac{1}{2} \times m \times (m-1)$ Boolean outcomes. The testing procedure for pair (i, j, u, v) is as follows. IT is investigated by checking whether $P(X_i = v, X_j = u | R_{(ij)} = r) \leq P(X_i = u, X_j = v | R_{(ij)} = r)$. If the sample fractions are reversely ordered (i.e., $\hat{P}(X_i = v, X_j = u | R_{(ij)} = r) > \hat{P}(X_i = u, X_j = v | R_{(ij)} = r)$, the McNemar (1947) test is used to decide whether the violation is significant. Let $n_{uv|r}$ and $n_{vu|r}$ denote the sample sizes of the relevant fractions, then under the null-hypothesis that the two bivariate conditional probabilities are equal,

$$X^2 = \frac{(n_{uv|r} - n_{vu|r})^2}{n_{uv|r} + n_{vu|r}}$$

has an asymptotic chi-square distribution with one degree of freedom. As with MS-CPM, violations smaller than 0.03 are ignored.

For the MS-CPM and IT methods, the backward item-selection procedures are formally identical to that of the MIIO method, and are not repeated here. For confirmatory results from the MS-CPM method, we infer that the LS-CPM supports the final item subset; and for confirmatory results from the IT method, we infer that the LS-ACM supports the final item subset. Many different strategies for testing the IT property are possible (see Sijtsma & Junker, 1996, p. 90) but they are beyond the scope of this study.

6. Real-Data Examples

We discuss two real-data examples to illustrate how the MIIO, MS-CPM, and IT methods may be used to investigate the latent scales. We used the function `check.iio` from the R package `mokken` (Van der Ark, 2007).

6.1. Ordering Coping Strategies

Data came from eight polytomous items administered to 828 respondents (Cavalini, 1992) asking them how they coped actively with the bad smell from a factory in the neighborhood of their homes. This was a survey research project, during which the items were tried for the first time. Table 1 shows the items ordered and numbered by increasing item mean. Items have four ordered answer categories, "never" (score 0), "seldom" (1), "often" (2), and "always" (3) (i.e., $m = 3$). The items constitute an ordinal scale according to the monotone homogeneity model (Sijtsma & Molenaar, 2002, Chapter 3). The items require global assessment using a rating scale (Van Engelenburg, 1997); hence, the LS-CPM may be the appropriate model to analyze the data. The aim of the analysis was to select a subset of items that constitute an LS-CPM scale, and represent a set of invariantly ordered coping reactions.

First, we tested the data for MIIO (Equation (18)), which is the least restrictive manifest ordering property, to identify items grossly violating an invariant ordering. We found that two out of the $\frac{1}{2} \times 8 \times 7 = 28$ item pairs (i.e., item pairs (5,6) and (5,7)) violated MIIO. Table 1 (Step 1) shows that item 5 was included in two item pairs violating MIIO, and items 6 and 7 were each included in one item pair. Removal of item 5 from the eight-item set (Table 1, Step 2) resulted in a seven-item set without violations, which provided support for IIO.

TABLE 1.
Violations of MIIO, MS-CPM, and IT for coping-strategy data.

Item Nr.	Mean	Wording	MIIO		MS-CPM		IT	
			Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
1.	0.264	Call environmental agency	0	0	2	0	2	NA
2.	0.353	File a complaint with producer	0	0	1	0	0	0
3.	0.535	Go elsewhere for fresh air	0	0	4	NA	NA	NA
4.	0.651	Experience unrest	0	0	2	0	1	0
5.	0.818	Try to find solutions	2	NA	NA	NA	NA	NA
6.	0.860	Do something to get rid of it	1	0	3	NA	NA	NA
7.	0.983	Talk to friends and family	1	0	2	0	1	0
8.	1.849	Search for source of malodor	0	0	0	0	0	0

Note: NA = Not available.

Next, we tested the MS-CPM (Equation (17)) for the remaining seven items, and found that seven out of the $\frac{1}{2} \times 7 \times 6 \times 3 = 63$ pairs of manifest ISRFs showed violations (Table 1; Step 3). Items 3 and 6 together were involved in all violations. Removal of these items from the seven-item set (Table 1, Step 4) resulted in a five-item scale for which the MS-CPM could not be rejected, and which provided support for the LS-CPM.

For the purpose of illustration, we also investigated the IT property (Equation (24)) for the remaining five items. Two out of the $\frac{1}{2} \times 5 \times 4 \times \frac{1}{2} \times 4 \times 3 = 60$ pairs of bivariate conditional probabilities violated IT; that is, $\hat{P}(X_1 = 2, X_4 = 3 | R_{(ij)} \in \{6, \ldots, 9\}) > \hat{P}(X_1 = 3, X_4 = 2 | R_{(ij)} \in \{6, \ldots, 9\})$ and $\hat{P}(X_1 = 2, X_7 = 3 | R_{(ij)} \in \{6, \ldots, 9\}) > \hat{P}(X_1 = 3, X_7 = 2 | R_{(ij)} \in \{6, \ldots, 9\})$ (Table 1, Step 5). Both were significant. Removal of item 1 from the five-item set resulted in a scale without violations (Table 1, Step 6), thus providing support for the LS-ACM. Because both violations were due to the same five respondents who had atypical item-score patterns, one may also argue that these respondents are outliers, and should be removed from the analysis. This was not pursued here.

6.2. *Dutch History*

The data were scores on three items collected from 752 students. The items were selected from a 40-item exam on Dutch history to illustrate the LS-ACM rather than the LS-CPM used in the first example. In each of the items, four historical events are presented and the student is asked whether the first event preceded the second, the second the third, and the third the fourth. The remaining 37 items had different item formats and could not be used for illustrating the LS-ACM.

The events of zero or one correct answer were relatively rare. Hence, the three items were scored as follows: 0 for zero or one correct answer; 1 for two correct answers; and 2 for three correct answers. Items were numbered following their ascending sample means, $\overline{X}_1 = 1.243$, $\overline{X}_2 = 1.327$, and $\overline{X}_3 = 1.386$. The task structure suggests that the subtasks may be solved in an arbitrary order (Van Engelenburg, 1997). Thus, the LS-ACM may be the appropriate model for investigating the item ordering. With only three items, item selection is not of interest here. It may be noted that for three items, the rest score is the score on one item only.

The LS-ACM was investigated checking $\frac{1}{2} \times 3 \times 2 \times \frac{1}{2} \times 3 \times 2 = 9$ pairs of bivariate conditional probabilities (Equation (24)), one of which was significant: $\hat{P}(X_2 = 1, X_3 = 0 | X_1 = 0) > \hat{P}(X_2 = 0, X_3 = 1 | X_1 = 0)$: $X^2 = 3.90$, $df = 1$, $p = .048$. Based on the IT method, the LS-ACM should be rejected. For completeness, we also used the MS-CPM and MIIO meth-

ods for data analysis. For MS-CPM, we found that two out of the six pairs of manifest ISRFs violated the MS-CPM. For MIIO we did not find violations.

7. Discussion

IIO is important in several applications of tests and questionnaires. For dichotomous items, Mokken's double monotonicity model and its special case, the Rasch model, have the IIO property; and fitting models support IIO for the application envisaged. For polytomous items, only highly restrictive IRT models have IIO. In this study, we explored the development of nonparametric IRT models, latent scales, for short, that have IIO, and we proposed methods for investigating IIO in polytomous item response data. These were the methods MIIO, MS-CPM, and IT, and we illustrated their use by means of two small data sets. Future research has to produce an example of a real-life application, but in several research applications IIO is already pursued, sometimes using the terminology of cumulative or hierarchical scales.

For realistic test length, methods MIIO, MS-CPM, and IT may produce many detailed results. This large number may complicate drawing conclusions about the fit of a latent scale. Ligtvoet et al. (2010a) proposed a method for investigating MIIO that reduces large numbers of detailed results to one final outcome. We adapted their method to the investigation of MS-CPM and IT and made the resulting data analysis much simpler and effective.

Rather than using local tests, one may prefer a global goodness-of-fit statistic, which assesses all violations simultaneously (but see Molenaar, 2004, who warns against the lack of diagnostic information provided by such global test results). Van der Ark, Croon and Sijtsma (2008) used marginal models for simultaneously testing properties of nonparametric IRT models, and this approach may also prove viable in the context of latent scales.

We ignored small violations of the MIIO, MS-CPM, and IT properties but did not adjust the nominal Type-I error-rate for multiple testing, which is consistent with model-fit investigation in nonparametric IRT (Sijtsma & Molenaar, 2002). More research is needed to find the proper balance between pre-selecting ignorable sample violations and an adequate Type I error rate. Different choices can be made with respect to conditioning variable Y. For the investigation of IT, the number of items investigated simultaneously may be varied. These and other topics not mentioned are studied in future research.

Requiring large numbers of IRFs not to intersect—that is, requiring IIO—means asking a lot of the items, and may easily lead to a loss of items resulting in an unacceptable reduction of total-score reliability. A solution to this problem may be identifying clusters of adjacent items in the overall item ordering and investigating whether the mean IRFs per cluster intersect, thus pursuing an invariant cluster ordering. This would provide some leeway for the items in the same cluster, thus accepting noise in the item ordering within clusters while maintaining an overall item ordering that might provide a good approximation to IIO. The final item ordering administered to individuals then follows the established ordering of item clusters but item ordering within clusters is free. This is yet another topic for future research.

Only a few studies have addressed the ordering of polytomous items, let alone IIO. This study provides a step in the direction of the development of a sound psychometric theory for latent scales and IIO of polytomous items, and of data-analysis methods that can be used for investigating whether a latent scale or IIO holds.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.
Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Bleichrodt, N., Drenth, P.J.D., Zaal, J.N., & Resing, W.C.M. (1987). *Revisie Amsterdamse kinder intelligentie test. Handleiding (Revision Amsterdam child intelligence test)*. Lisse, The Netherlands: Swets & Zeitlinger.
Cavalini, P.M. (1992). *It's an ill wind that brings no good. Studies on odour annoyance and the dispersion of odorant concentrations from industries.* Unpublished doctoral dissertation, University of Groningen, The Netherlands.
Chang, H., & Mazzeo, J. (1994). The unique correspondence of the item response function and item category response function in polytomously scored item response models. *Psychometrika*, *59*, 391–404.
Douglas, R., Fienberg, S.E., Lee, M.-L.T., Sampson, A.R., & Whitaker, L.R. (1991). Positive dependence concepts for ordinal contingency tables. In H.W. Block, A.R. Sampson, & T.H. Savits (Eds.), *Topics in statistical dependence* (pp. 189–202). Hayward, CA: Institute of Mathematical Statistics.
Emons, W.H.M., Sijtsma, K., & Meijer, R.R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*, 105–120.
Glas, C.A.W., & Verhelst, N.D. (1995). Testing the Rasch model. In G.H. Fischer, & I.W. Molenaar (Eds.), *Rasch models: foundations, recent developments, and applications* (pp. 69–96). New York: Springer.
Hemker, B.T., Sijtsma, K., Molenaar, I.W., & Junker, B.W. (1997). Stochastic ordering using the latent trait and the sum score in polytomous IRT models. *Psychometrika*, *62*, 331–347.
Hemker, B.T., Van der Ark, L.A., & Sijtsma, K. (2001). On measurement properties of continuation ratio models. *Psychometrika*, *66*, 487–506.
Hollander, M., Proschan, F., & Sethuraman, J. (1977). Functions decreasing in transposition and their applications in ranking problems. *The Annals of Statistics*, *5*, 722–733.
Jansen, B.R.J., & Van der Maas, H.L.J. (1997). Statistical test of the rule assessment methodology by latent class analysis. *Developmental Review*, *17*, 321–357.
Ligtvoet, R., Van der Ark, L.A., Te Marvelde, J.M., & Sijtsma, K. (2010a). Investigating an invariant item ordering for polytomously scored items. *Educational and Psychological Measurement*, *70*, 578–595.
Ligtvoet, R., Van der Ark, L.A., Bergsma, W.P., & Sijtsma, K. (2010b). *Examples concerning the relationships between latent/manifest scales* (unpublished manuscript). Retrieved from http://spitswww.uvt.nl/~avdrark/research/LABSexamples.pdf.
Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.
McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, 153–157.
Mellenbergh, G.J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*, 91–100.
Mokken, R.J. (1971). *A theory and procedure of scale analysis*. The Hague/Berlin: Mouton/De Gruyter.
Molenaar, I.W. (1983). *Item steps (Heymans Bulletin 83-630-EX)*. Groningen, The Netherlands: University of Groningen, Department of Statistics and Measurement Theory.
Molenaar, I.W. (1997). Nonparametric models for polytomous responses. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 369–380). New York: Springer.
Molenaar, I.W. (2004). About handy, handmade and handsome models. *Statistica Neerlandica*, *58*, 1–20.
Molenaar, I.W., & Sijtsma, K. (2000). *User's Manual MSP5 for Windows*. Groningen, The Netherlands: iec ProGAMMA.
Muraki, E. (1990). Fitting a polytomous item response model to Likert-type data. *Applied Psychological Measurement*, *14*, 59–71.
Muraki, E. (1992). A generalized partial credit model: applications for an EM algorithm. *Applied Psychological Measurement*, *16*, 159–177.
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Nielsen and Lydiche.
Rosenbaum, P.R. (1987a). Probability inequalities for latent scales. *British Journal of Mathematical & Statistical Psychology*, *40*, 157–168.
Rosenbaum, P.R. (1987b). Comparing item characteristic curves. *Psychometrika*, *52*, 217–233.
Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph* (No. 17).
Samejima, F. (1997). Graded response model. In W.J. van der Linden, & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.
Scheiblechner, H. (1995). Isotonic ordinal probabilistic models (ISOP). *Psychometrika*, *60*, 281–304.
Scheiblechner, H. (2003). Nonparametric IRT: testing the bi-isotonicity of Isotonic Probabilistic Models (ISOP). *Psychometrika*, *68*, 79–96.
Shaked, M., & Shantikumar, J.G. (1994). *Stochastic orders and their applications*. San Diego, CA: Academic Press.
Sijtsma, K., & Hemker, B.T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*, *63*, 183–200.
Sijtsma, K., & Junker, B.W. (1996). A survey of theory and methods of invariant item ordering. *British Journal of Mathematical & Statistical Psychology*, *49*, 79–105.
Sijtsma, K., Meijer, R.R., & Van der Ark, L.A. (2011). Mokken Scale Analysis as time goes by: an update for scaling practitioners. *Personality and Individual Differences*, *50*, 31–37.
Sijtsma, K., & Molenaar, I.W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage.
Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical & Statistical Psychology*, *43*, 39–55.
Van der Ark, L.A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, *20*(11), 1–19.

Van der Ark, L.A., Croon, M.A., & Sijtsma, K. (2008). Mokken scale analysis for dichotomous items using marginal models. *Psychometrika*, *73*, 183–208.

Van der Ark, L.A., Hemker, B.T., & Sijtsma, K. (2002). Hierarchically related nonparametric IRT models, and practical data analysis methods. In G.A. Marcoulides, & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 41–62). Mahwah, NJ: Erlbaum.

Van Engelenburg, G. (1997). *On psychometric models for polytomous items with ordered categories within the framework of item response theory*. Unpublished doctoral dissertation, University of Amsterdam.

Van Schuur, W.H. (2003). Mokken scale analysis: between the Guttman scale and parametric item response theory. *Political Analysis*, *11*, 139–163.

Watson, R., Deary, I.J., & Shipley, B. (2008). A hierarchy of distress: Mokken scaling of the GHQ-30. *Psychological Medicine*, *38*, 575–579.

Wechsler, D. (2003). *Wechsler intelligence scale for children* (4th ed.). San Antonio, TX: The Psychological Corporation.

Weekers, A.M., Brown, G.T.L., & Veldkamp, B.P. (2009). Analyzing the dimensionality of the Student's Conceptions of Assessment inventory. In D.M. McInerney, G.T.L. Brown, & G.A.D. Liem (Eds.), *Student perspectives on assessment: what students can tell us about assessment for learning* Charlotte, NC: Information Age.

Manuscript Received: 6 DEC 2009
Final Version Received: 13 SEP 2010
Published Online Date: 27 JAN 2011